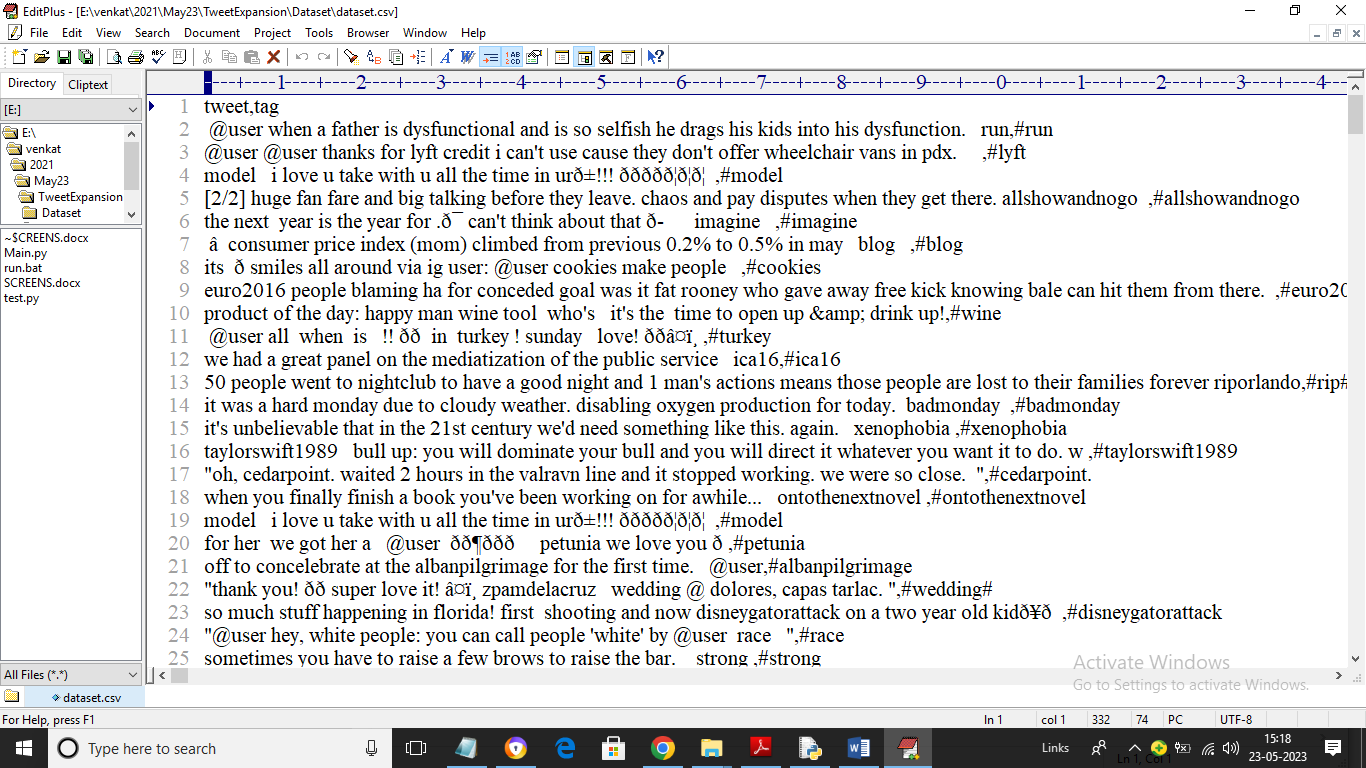
Hashtag-Based Tweet Expansion for Improved Topic Modelling

In this paper author employing BILSTM, BERT and GRAPHCNN algorithms to predict topics from heterogeneous and homogenous tweets where heterogeneous refers to unrelated tweets and homogenous refers to related tweets. To predict topics from text many existing algorithms are available such as tweet aggregation, where related tweets are combined into a single document or 2) tweet expansion with related text from external sources. The first approach faces the problem of losing the topic distribution in individual tweets. While finding a relevant text from the external source for a random tweet in the second approach is challenging for various reasons like differences in writing styles, multilingual content, and informal text. In contrast to adding context from external resources or combining related tweets into a pool, this study uses the internal vocabulary (hashtags) to counter under-specificity and sparsity in tweets. Earlier studies have indicated hashtags to be an important feature for representing the underlying context present in the tweet. Sequential models like Bi-directional Long Short Term Memory (BiLSTM) and Convolution Neural Network (CNN) over distributed representation of words have shown promising results in capturing semantic relationships between words of a tweet in the past. Motivated by the above, this article proposes a unified framework of hashtag-based tweet expansion exploiting text-based and network-based representation learning methods such as BiLSTM, BERT, and Graph Convolution Network (GCN). The hashtag-based expanded tweets using the proposed framework have significantly improved topic modelling performance compared to un-expanded (raw) tweets and hashtag-pooling-based approaches over two real-world tweet datasets of different nature.

Propose paper will expand or identify TWEETS from given Hashtags and then identify topics from expanded Hashtags

BILSTM, GRAPHCNN and BERT are based on Text processing which can be used to efficiently process and predict topics from the tweets. It’s difficult to train all 3 algorithms so we are training BERT and BILSTM

To train above algorithms author is using tweets from various HASHTAGS and not publish those tweets on internet so we have collected some tweets from various hashtags and below are the dataset details.



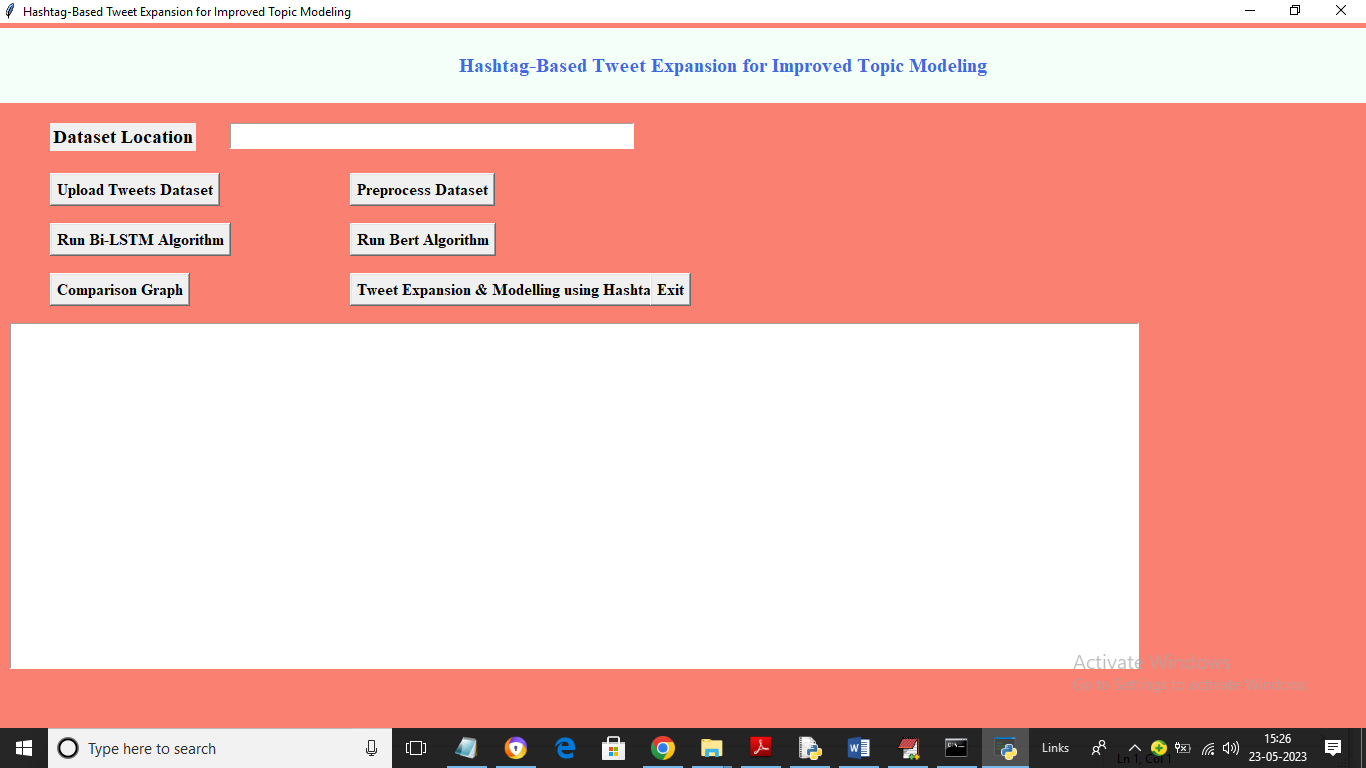
In above dataset screen first row contains dataset column names and remaining rows contains dataset values and by using above tweets and tags will train BILSTM and BERT algorithm and then evaluate performance using FSCORE and PRECISION

To implement this project we have designed following modules

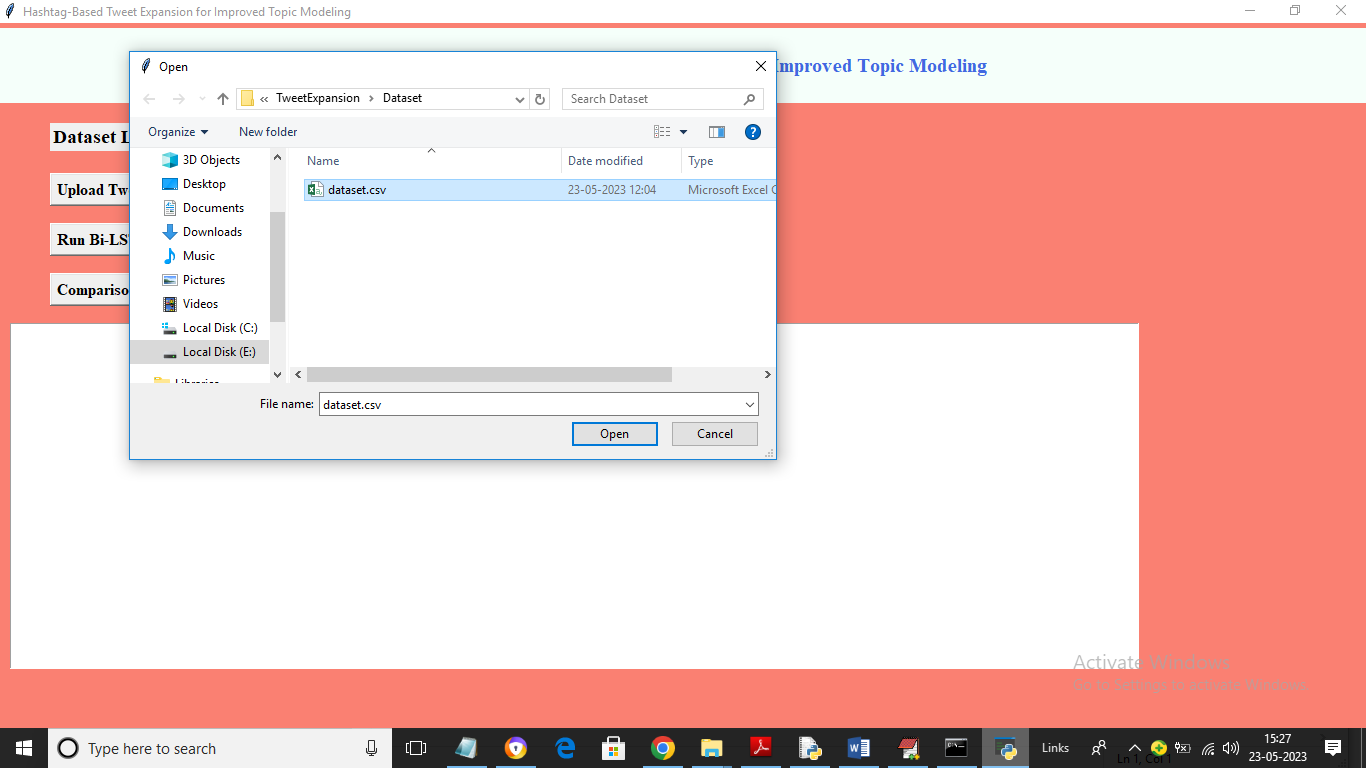
1. Upload Tweets Dataset: using this module we will upload dataset to application
2. Preprocess Dataset: using this module we will read each tweet from dataset and then remove stop words, special symbols and then apply stemming and then extract TOP K WORDS
3. Run Bi-LSTM Algorithm: using this module we will convert all tweets into TFIDF numeric vector which will replace each words with its average frequency and then input to BILSTM algorithm to train and model and this model will be applied on test data to calculate precision and FSCORE
4. Run Bert Algorithm: using this module we will convert all tweets into TFIDF numeric vector which will replace each words with its average frequency and then input to BERT algorithm to train and model and this model will be applied on test data to calculate precision and FSCORE
5. Comparison Graph: using this module we will plot precision and FCSORE graph between both algorithms
6. Tweet Expansion & Modelling using Hashtag: using this module we will input HASHTAG words and then algorithm will expand HASHTAG to tweet and then extract topics

SCREEN SHOTS

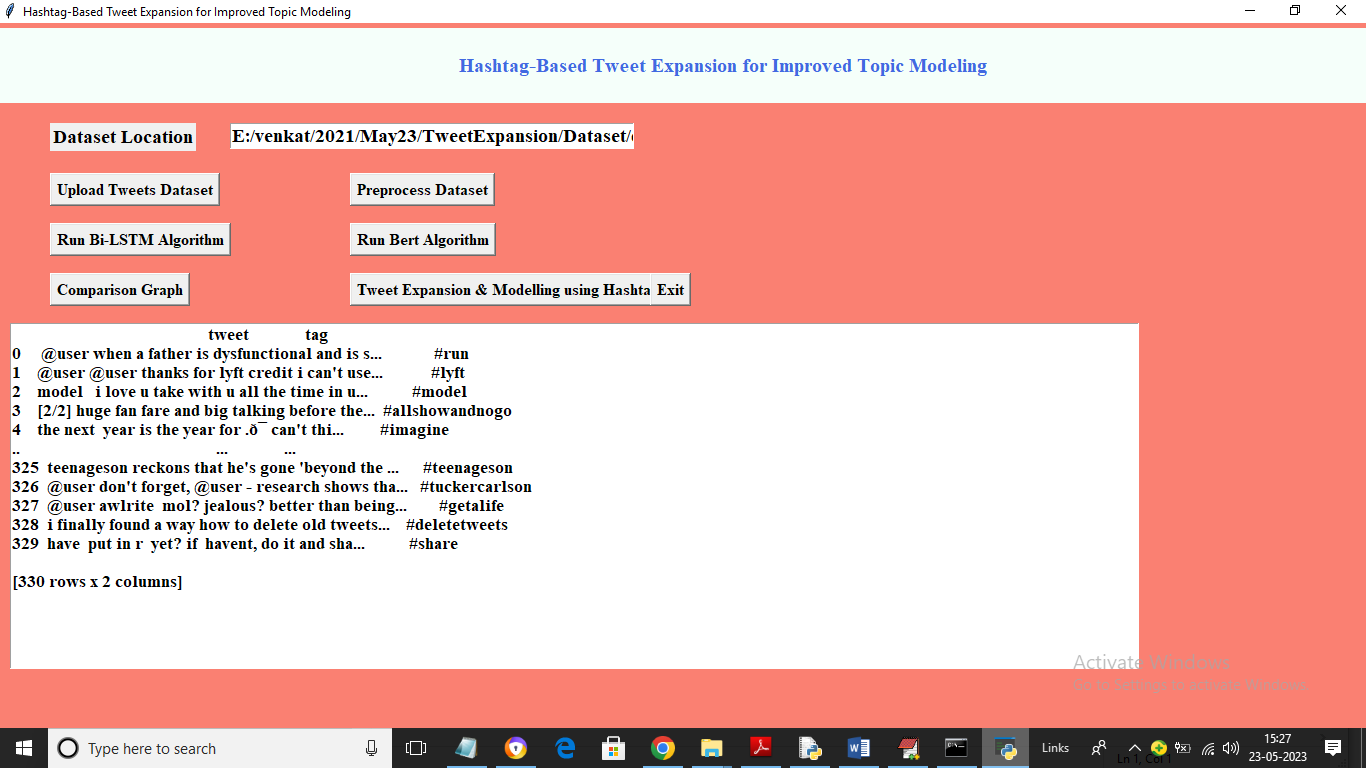
To run project double click on ‘run.bat’ file to get below screen



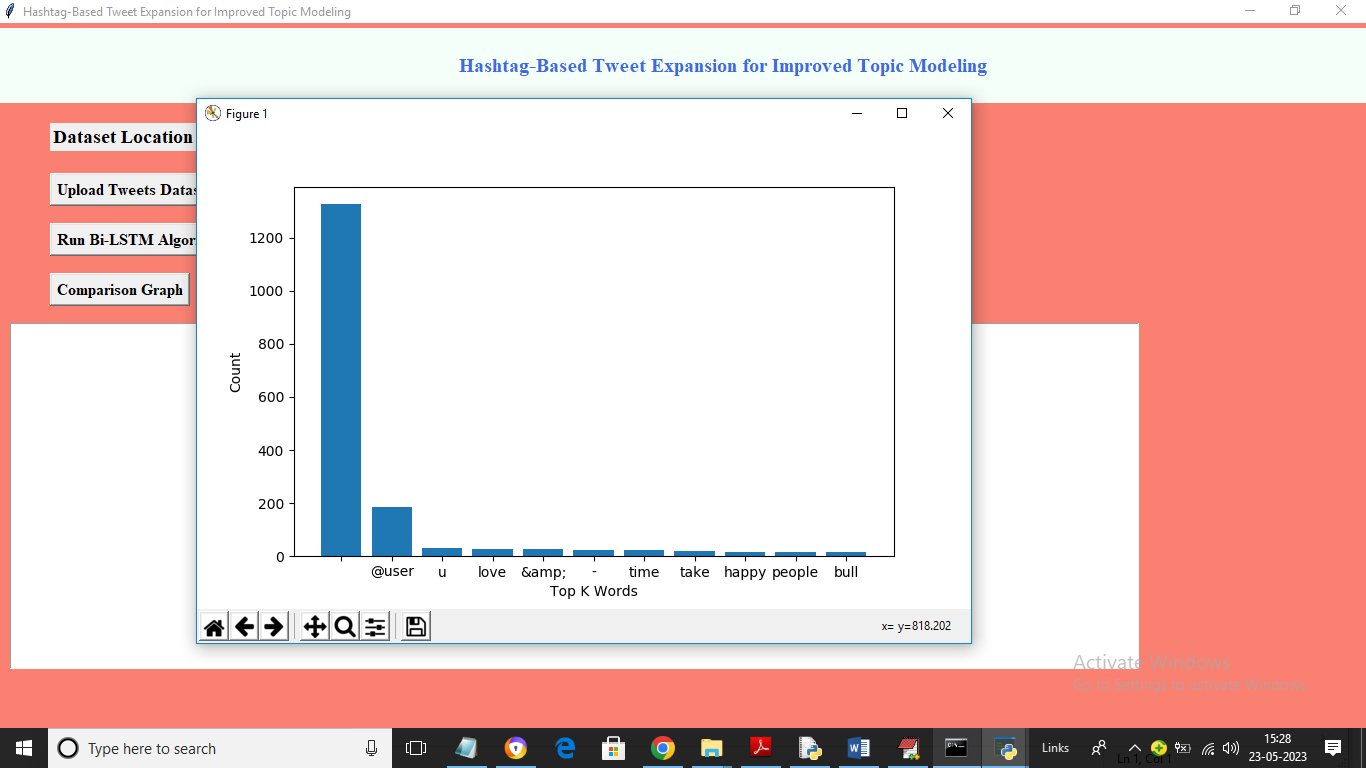
In above screen click on ‘Upload Tweets Dataset’ button to load tweets and get below output



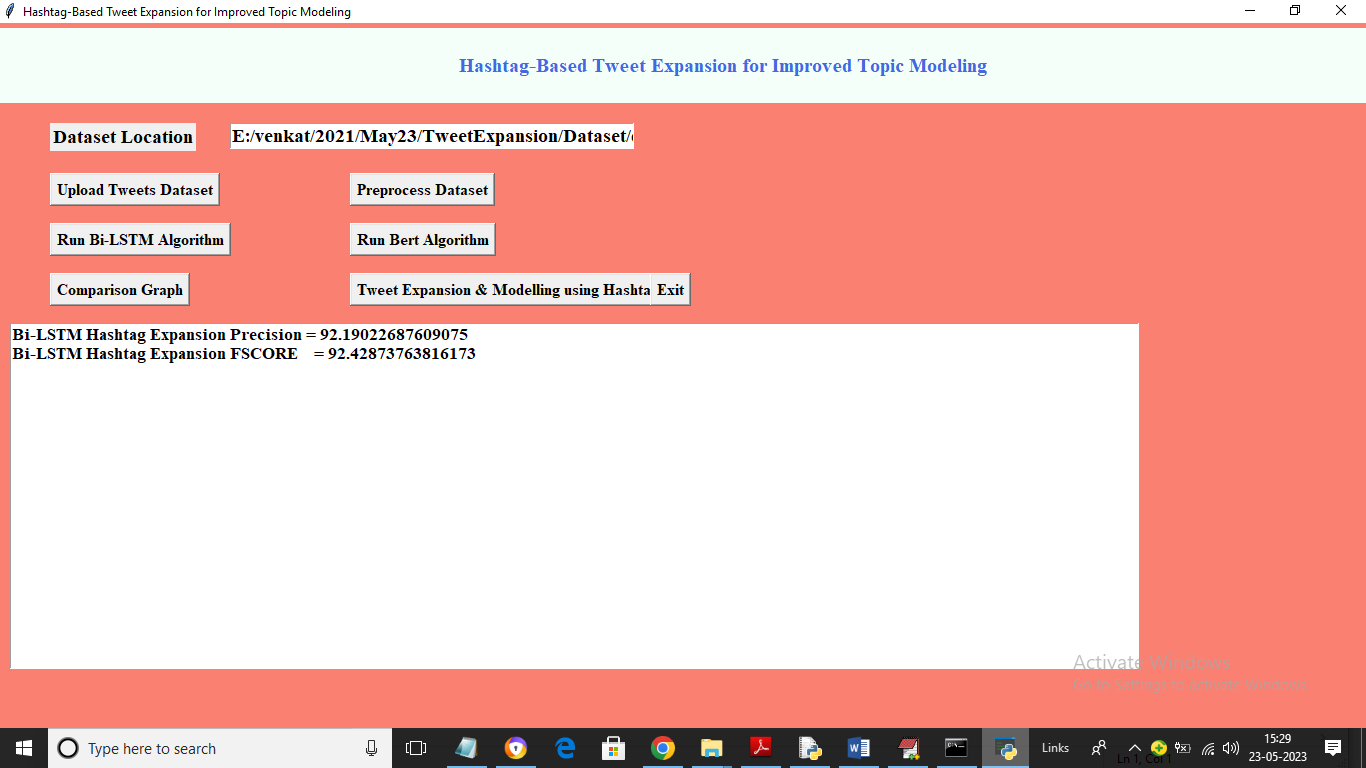
In above screen selecting and uploading ‘dataset.csv’ file and then click on ‘Open’ button to load dataset and get below screen



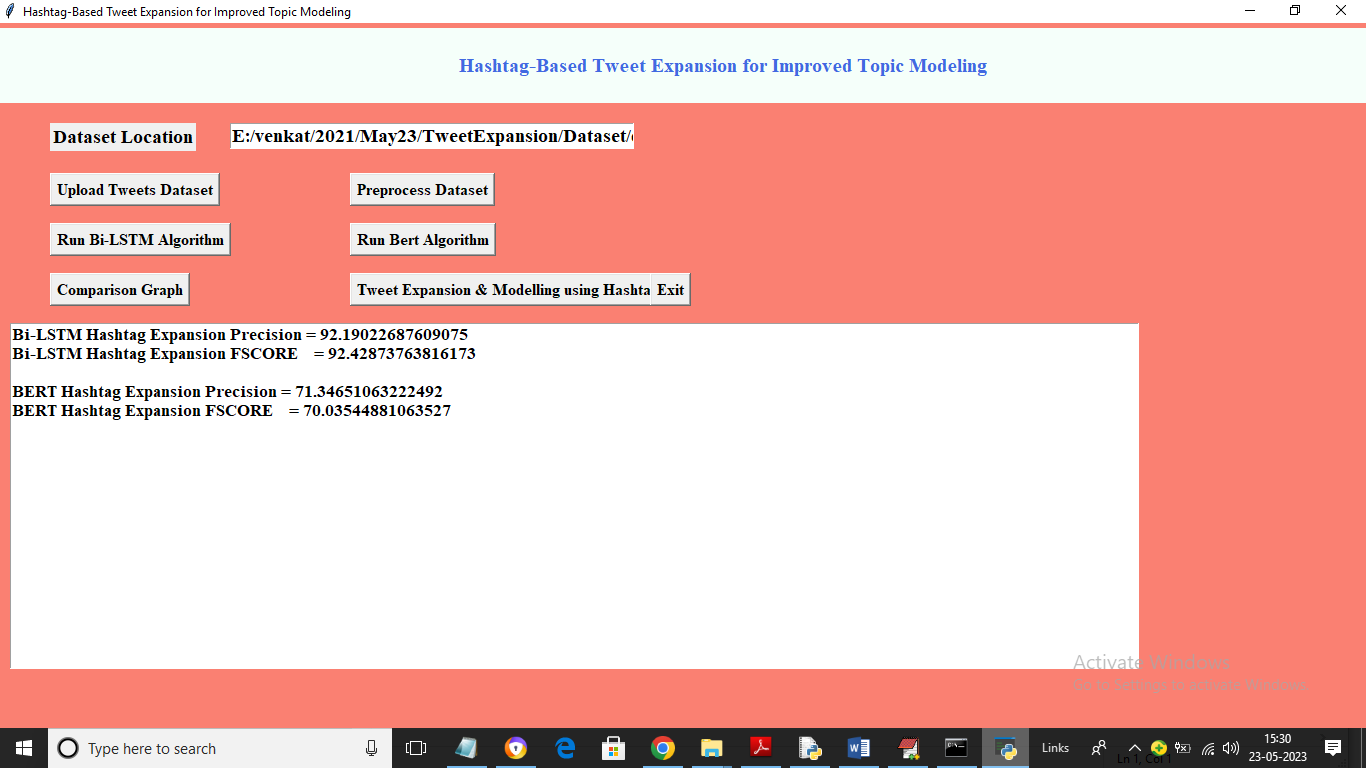
In above screen dataset loaded and now click on ‘Preprocess Dataset’ button to process tweets and then find and plot TOP K WORDS graph



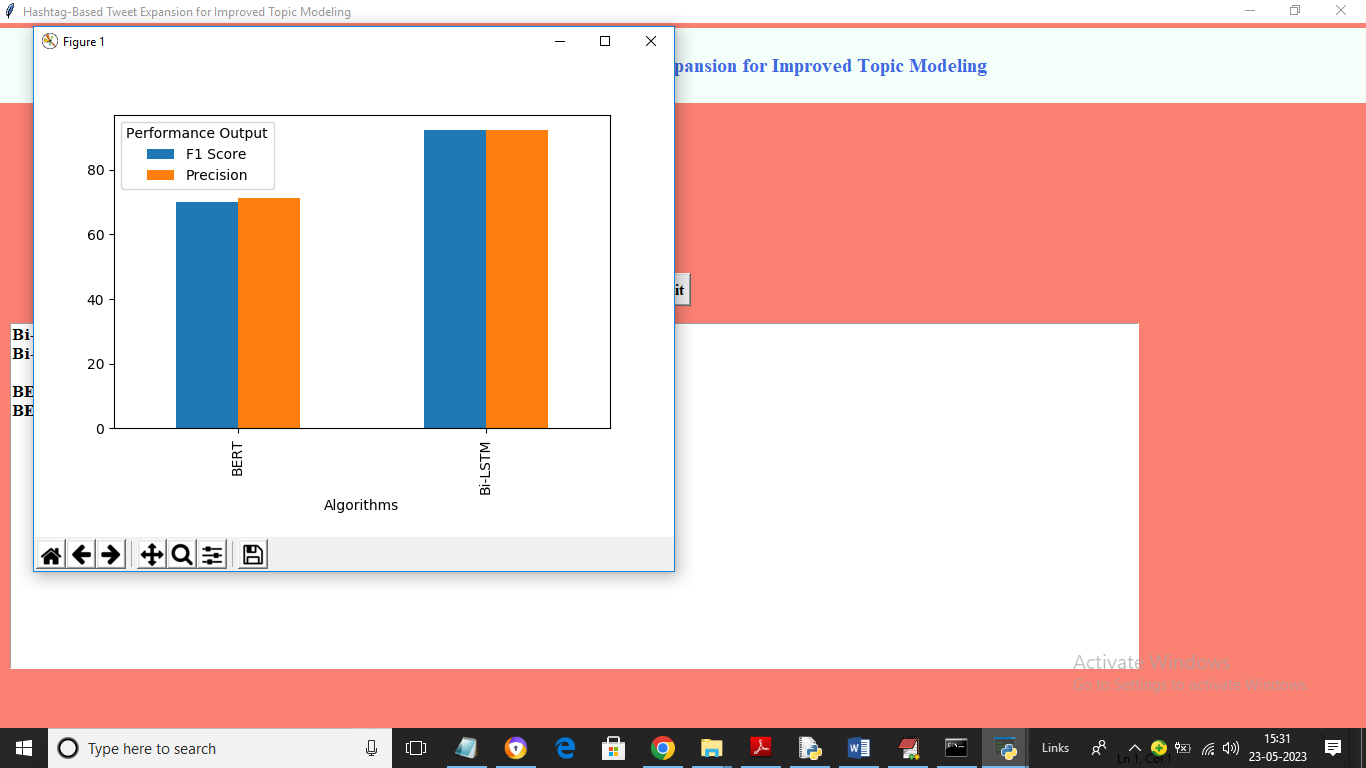
In above screen dataset processing completed and graph showing TOP K WORDS graph where x-axis represents words and y-axis represents frequency of words. Now click on ‘Run Bi-LSTM Algorithm’ button to train BILSTM and get below output



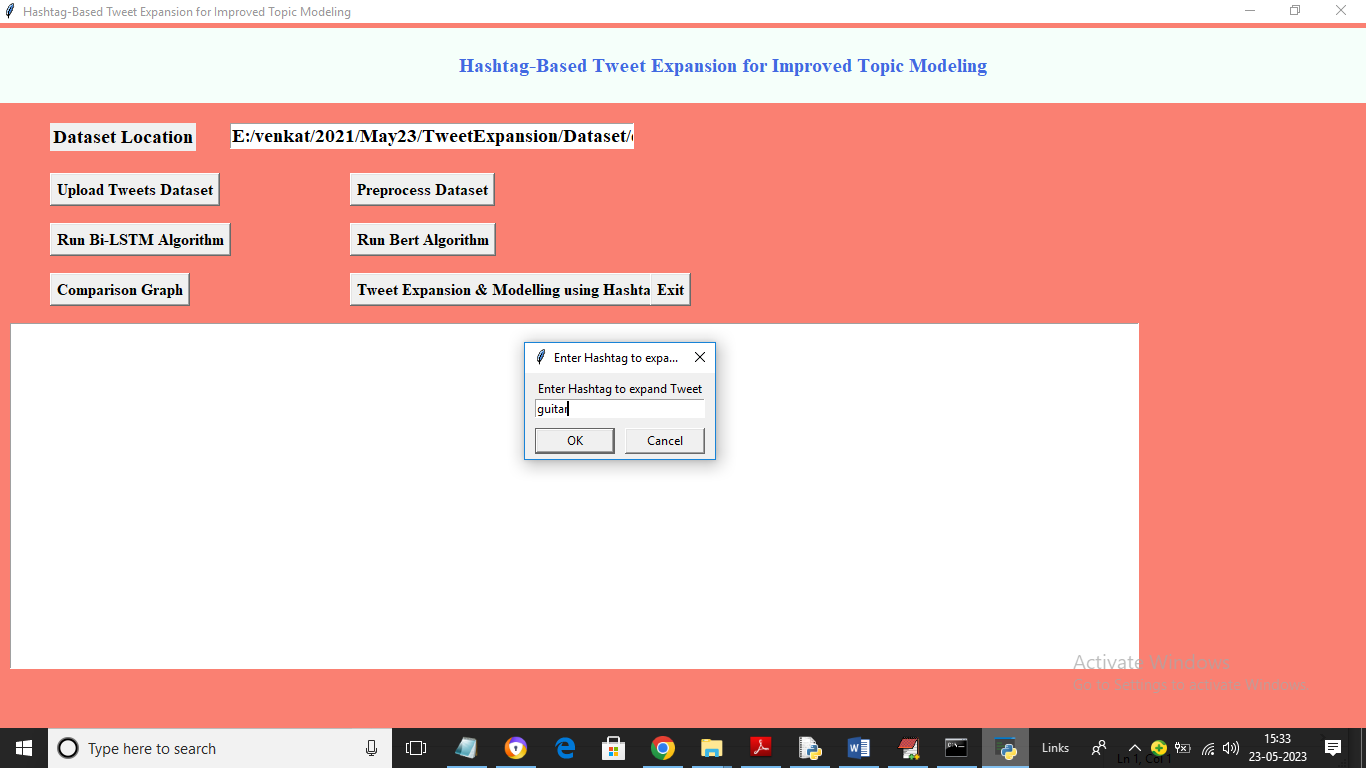
In above screen BILSTM training completed and with BILSTM we got 92% precision and FSCORE value and now click on ‘Run Bert Algorithm’ button to train BERT and get below output



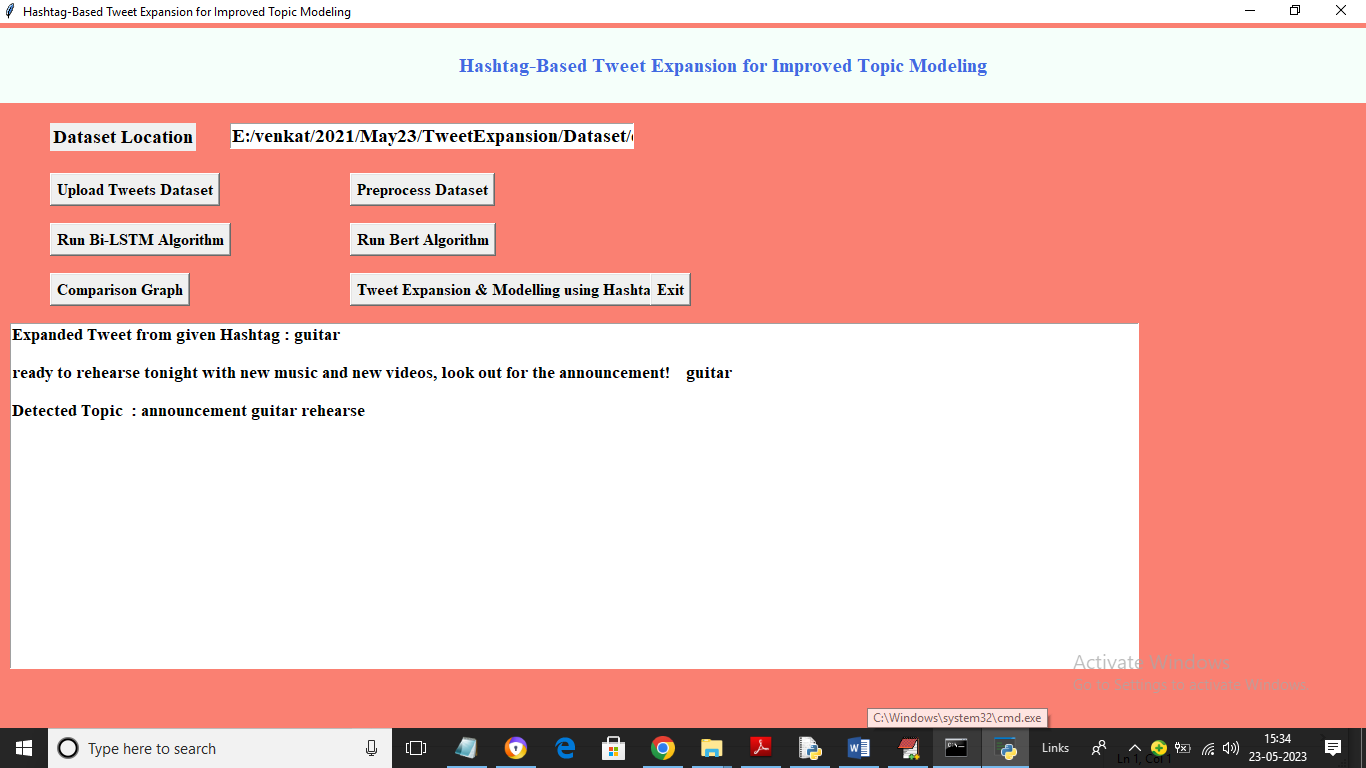
In above screen with BERT algorithm we got FSCORE and precision as 71% and now click on ‘Comparison Graph’ button to get below graph



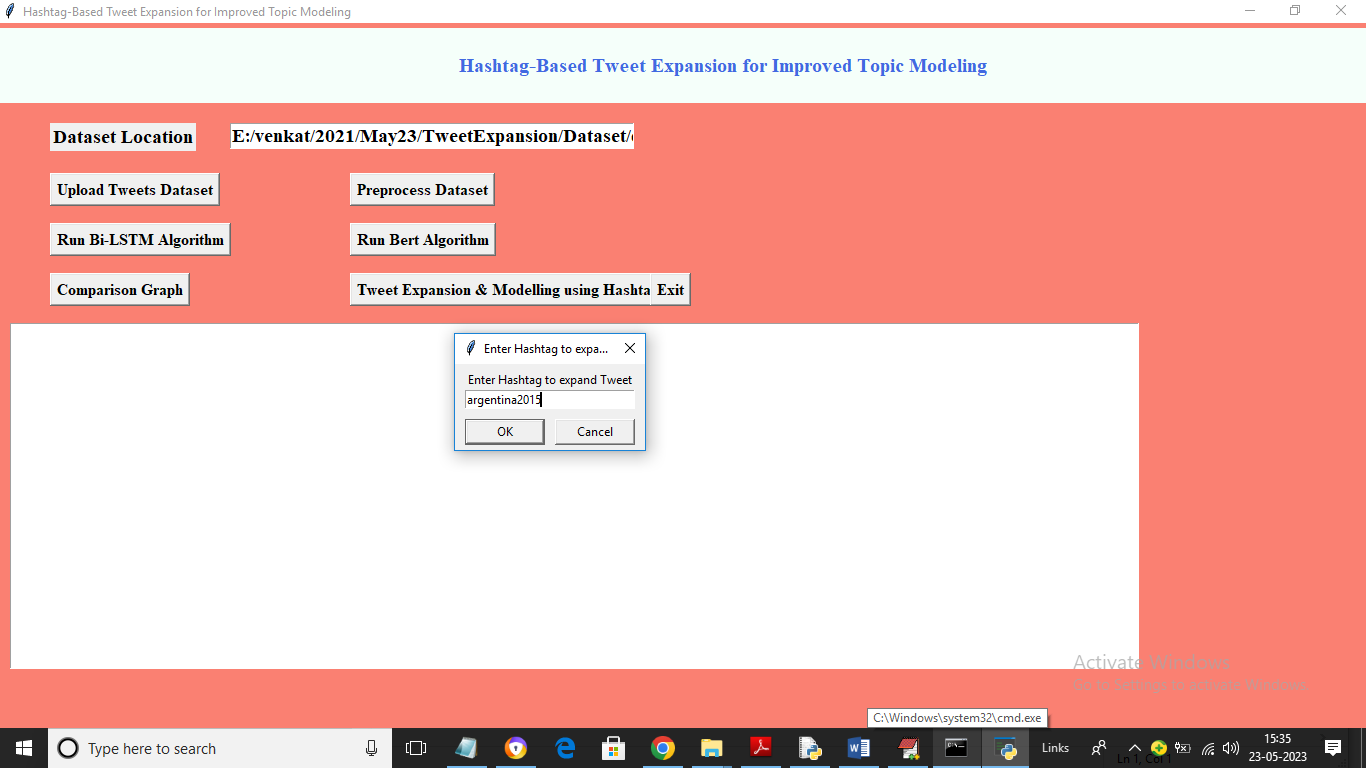
In above graph x-axis represents algorithm names and y-axis represents precision and FSCORE in different colour bars and in both algorithms BILSTM has got better result. Now click on ‘Tweet Expansion & Modelling using Hashtag’ to input hashtag and get expanded tweets and topic



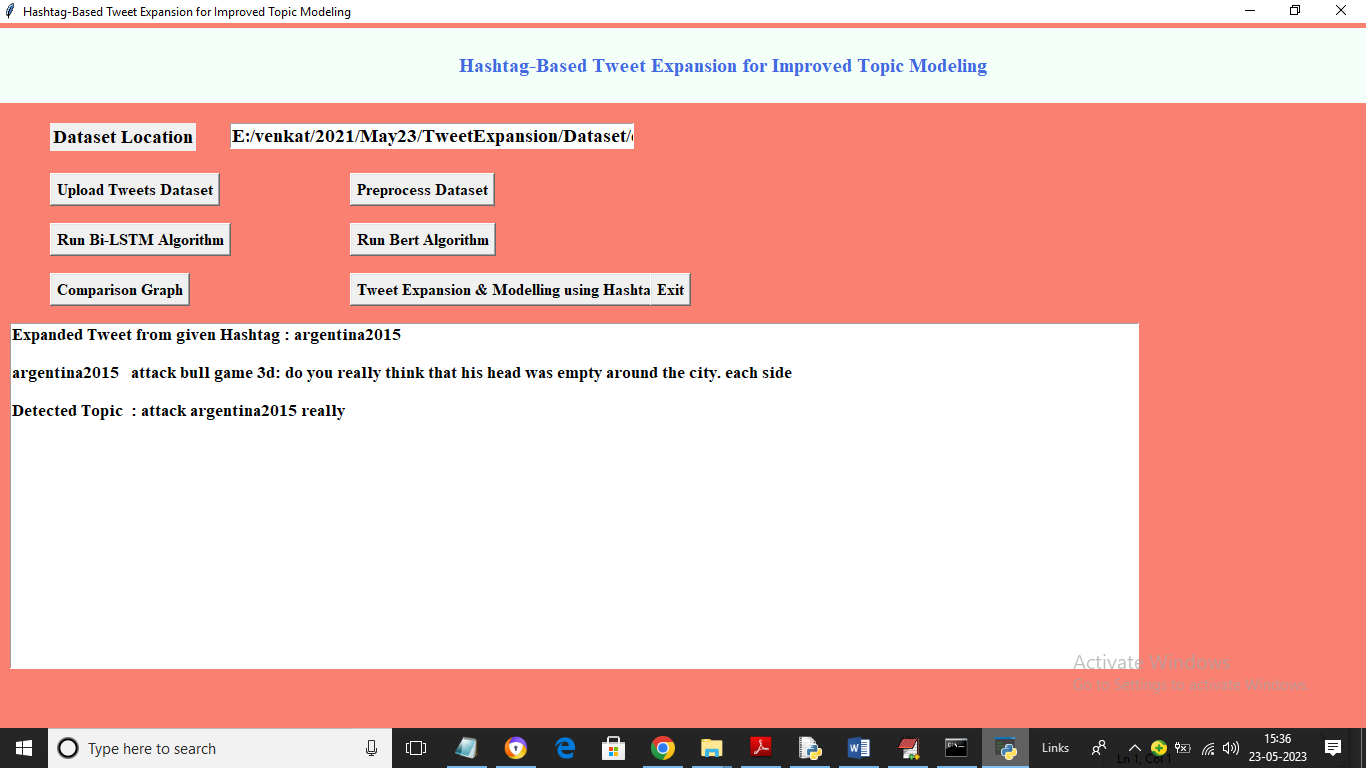
In above screen in dialog box I entered Hashtag as ‘guitar’ and then click on ‘OK’ button to get below output



In above screen for given HASHTAG we got expanded TWEET and then displaying TOPICS as ‘announcement guitar rehearse ‘. Similarly enter any Hashtag from dataset and get expanded tweets and TOPICS



In above screen I entered some Hashtag and then click OK to get below output



In above screen for given Hashtag we got expanded tweets and TOPICS